AN EVALUATION OF THE FAIRNESS OF THE FLIGHT APTITUDE SELECTION TEST (FAST)

John A. Dohme, Ph.D.
Research Psychologist
US Army Research Institute
Fort Rucker Field Unit

The concept "test fairness" has developed only recently. A major impetus in the development and application of the concept has come from the publication of the Uniform Guidelines on Employee Selection Procedures (UGES) in 1978. The UGES are interpreted as mandating the use of a regression model in evaluating test fairness. A technique was developed utilizing a regression model to evaluate the fairness of the Flight Aptitude Selection Test (FAST) for the groups identified by the UGES: Blacks, American Indians, Asians, Hispanics, Caucasians and females. The regression of FAST scores on overall grades in the Initial Entry Rotary Wing (IERW) course was performed for each of the above groups in comparison with the majority group. Available population sizes were considered too small to permit a conclusive fairness evaluation at this time. The fairness evaluation will be repeated semiannually until minority population sizes permit sufficient power to perform a definitive analysis.

AN EVALUATION OF THE FAIRNESS OF THE FLIGHT APTITUDE SELECTION TEST (FAST)

"Fairness" as a criterion for the evaluation of a test or other selection procedure is a relatively new concept. The concept has evolved from the technology of test validation to answer the question, "Is this test/procedure valid for the selection of minority as well as majority applicants?" Appropriate methodology for the evaluation of fairness is currently a matter for debate in the technical literature (Ledvinka, 1979). A major impetus for the development of fairness methodologies was the publication of Guidelines on Employee Selection Procedures in 1970 by the Equal Employment Opportunity Commission. In fact, the most current version, the Uniform Guidelines on Employee Selection (UGES) (1978), noted that, "The concept of fairness or unfairness of selection procedures is a developing concept, (14B(8))." Since this technology is still developmental, this paper will review the rationale and precedence for the FAST fairness evaluation in some detail.

Technical standards for performing a fairness evaluation are addressed by both professional and government agencies. The American Psychological Association (APA) publication, Principles for the Validation and Use of Personnel Selection Procedures (1975), discusses both technical and ethical implications of the choice of methodology in fairness research designs. The government publication referenced above, Uniform Guidelines on Employee Selection Procedures (UGES) published in 1978, which is a codified position agreed upon by the US Civil Service Commission, the Department of Justice, the EEOC, and the Department of Labor falls under the scope of Title VII of the 1964 Civil Rights Act

and, for that reason, carries the impact of law.[1] Furthermore, the current version of the UGES was reviewed by the APA prior to publication, thus, it is a synthesis of professional and governmental guidance in the technical and ethnical and legal aspects of fairness research designs. For these reasons, this paper will make frequent reference to the UGES.

The UGES define fairness by stating its obverse: "When members of one race, sex, or ethnic group characteristically obtain lower scores on a selection procedure than members of another group, and the differences in scores are not reflected in differences in a measure of job performance, use of the selection procedure may unfairly deny opportunities to members of the group that obtains the lower scores (Section 14B8a)." This definition has clear implications in the design of a fairness research study in that it specifies that fairness should be defined in terms of the bivariate distribution of test (or other selection procedure) scores and job performance scores. Specifically, fairness is demonstrated by coincident regression of job performance scores on test scores for a minority group and the majority group. Fairness does not require that minority performance on the test, or on the job be equal to majority performance but only that the test (or selection procedure) does not over or under predict minority performance vis a vis majority performance.

The UGES do not require routine demonstration of the fairness of a selection procedure for every minority group identified in section 4B (Blacks, American Indians, Asians, Hispanic and Caucasians). Section 14B(8)(b) states: "Where a selection procedure results in an adverse impact on a race, sex, or ethnic group identified in accordance with the classifications set forth in section 4 above and that group is a significant factor in the relevant labor market, the user generally should investigate the possible existence of unfairness for that group if it is technically feasible to do so." In other words, a demonstration of fairness is required only where:

(1) there is evidence of adverse impact as defined in section 4D of the UGES;

(2) that adverse impact affects a group identified in section 4B of the UGES;

(3) the group(s) affected comprise a significant factor in the relevant labor market which is defined in section 15A(1)(c) as constituting more than 2% of the labor force in a "relevant labor area";

(4) it is "technically feasible" to investigate the fairness issue. Technical feasibility is defined in section 14B(8)(c) to include:

(a) sufficient sample sizes to achieve statistical significance;

(b) direct comparability of the samples in terms of the actual jobs performed.

---

[1]At this writing, military personnel in DOD agencies do not fall under the purview of Title VII, thus, may not be legally bound to the UGES. However, the author takes the position that the UGES represent current professional thinking in this technical area, therefore, they provide appropriate guidance independent of their status as law.

The issues raised in paragraphs 1-3 above are empirical questions. They are best answered by descriptive data pertaining to the population of applicants to US Army flight training. The Fort Rucker Field Unit of ARI began an investigation of the selection rates of applicants of the groups identified in section 4B of the UGES. Data were requested from MILPERCEN and RCPAC and a quality check was performed on the data obtained from the master files. Master file data were cross referenced with data in the student pilot's flight folders at the Directorate of Training at Fort Rucker. Taking the black group as an example, master file data were missing for over 78% of the trainees, i.e., 78% of individuals who had entered the flight training course did not appear in the master file. Therefore, it must be concluded that the selection rates prior to 1980 are indeterminate and adverse impact cannot be assessed.

With the advent of the revised FAST test (RFAST) which replaced the earlier form in the field in early 1980, the data collection problem referenced above has been alleviated. The RFAST answer sheet requests information on the sex and ethnic status of applicants. All RFAST answer sheets are sent to ARI, Fort Rucker for machine scoring and storage in the RFAST archives, thus, all the information needed to determine whether or not adverse impact exists will be available at ARI Fort Rucker. Given that it commonly takes more than one year between taking the RFAST and graduation from the 34 week training program, it will be some time before adverse impact can be determined for the RFAST.

In the interim, the conservative assumptions will be made that adverse impact does exist for all the groups identified by Section 4B of the UGES, and that each of those groups constitutes more than 2% of the applicant population. Pursuant to Section 14B(8) of the UGES, a fairness evaluation will be undertaken for each group where it is "technically feasible" to do so. However, the issue of technical feasibility is, like the issue of fairness, a matter of some debate in the technical literature. As noted above, the UGES discuss the issue of technical feasibility with reference to sample size and comparability. In an empirical study of the statistical power associated with various sample sizes, Schmidt, Hunter and Urry (1976) concluded:

> *"This study demonstrates that sample sizes required to produce adequate power in empirical validation studies are substantially larger than has typically been assumed. This finding leads to the conclusion that, from the viewpoint of sample-size requirements, criterion-related validity studies are "technically feasible" much less frequently than is commonly assumed (p. 473)."*

Using the methodology developed by Schmidt, Hunter and Urry (1976) to estimate the sample size required in the present evaluation, and making the liberal assumptions that (1) the true validity of the FAST test is .50, (2) the reliability of the Initial Entry Rotary Wing (IERW) overall grade is .60 and (3) 70% of the applicants to the IERW program are accepted, 128 subjects per group would be required to reach a power of .90 (i.e., to have a 90% probability of rejecting the null hypothesis if it is indeed false). Thus, from the standpoint of the Schmidt, Hunter and Urry (1976) article, it is not technically feasible to perform a fairness evaluation of the FAST until a larger sample of IERW graduates is available.

An earlier section of this research report noted that a revised version of the FAST (the RFAST) is presently being implemented in the field. The version of the FAST being evaluated for fairness in this report has two different forms developed for implementation with commissioned officers and enlisted personnel respectively. Since the two forms differ substantially in content and number of items, the current fairness evaluation must be conducted separately for these two populations. There is only one form of the RFAST which has been developed for use with both populations. Therefore, future fairness evaluations will not require separate commissioned and enlisted samples which will considerably ameliorate the problem of collecting samples large enough to permit a conclusive fairness evaluation.

One key issue in the design of a fairness evaluation study is the choice of a statistical model to guide the minority/majority comparisons. Section 14B8 of the UGES raises the point that the concept fairness is still evolving in the literature. Specifically, the choice of a statistical model has been debated for nearly a decade since the publication of the 1970 version of the EEOC Guidelines (see Cole, 1972; Hunter and Schmidt, 1974; Hunter, Schmidt and Rauschenberger, 1977 and Ledvinka, 1979). The current literature focuses on four models which lead to different operational definitions of fairness/unfairness:

1. The regression model (Cleary, 1968) which states that a test is fair if the regression lines predicting job performance are the same (plus or minus sampling variation) for minority and majority groups.

2. The conditional probability model (Darlington, 1971; Cole, 1973) which states that a test is fair if the probability of being selected is the same for minority and majority group members who are actually capable of satisfactory job performance.

3. The constant ratio model (Thorndike, 1971) which states that a test is fair if its selection ratio for minority and majority groups is the same as the selection ratio using a perfectly valid test (or using the criterion measure itself for selection).

4. The quota model which states that a test is fair if its selection ratio is the same for all minority and majority groups regardless of group performance on the job.

While various authors continue to argue the technical and ethnical merits of these models, it has been pointed out by Ledvinka (1979, p. 552) and by Hunter, Schmidt and Rauschenberger (1977, p. 256) that the UGES clearly specify the regression model as being legally appropriate in the conduct of fairness research. Two UGES passages can be cited to document this point.

> *"When members of one race, sex or ethnic group characteristically obtain lower scores on a selection procedure than members of another group, and the differences in scores are not reflected in differences in a measure of job performance . . . ." (Section 14B8a)."*

"If unfairness is demonstrated through a showing that members of a particular group perform better or poorer on the job, then their scores on the selection procedure would indicate through comparison with how members of other groups perform, the user may either revise or replace the selection instrument in accordance with these guidelines, or may continue to use the selection instrument operationally with appropriate revisions in its use to assure compatibility between the probability of successful job performance and the probability of being selected." (Section 24B8d).

There is an additional, independent reason to use the regression model in this fairness evaluation. Of the four models, it alone does not require a "pass through" methodology in which IERW applicants are selected for flight training regardless of their FAST scores. While a pass through methodology is technically appropriate in fairness research, it incurs a substantial increase in attrition rate over the use of an efficacious selection procedure. Given that the training costs in the IERW program exceed $125,000 per trainee, the two costs of a pass through program, higher attrition costs and a reduced output of trainees, could conceivably cost the government millions of dollars per year and lead to an even greater shortfall in aviators in the field.

METHOD

The subjects that comprise the minority/female samples include all IERW program trainees who identified themselves as belonging to one of the groups previously identified in the UGES (Black, Hispanic, Asian, American Indian, female) and for whom both FAST and IERW overall grade (OAG) data were available in US Army Aviation Center (USAAVNC) records. The data collected cover the time span July 1975 to July 1979.

In order to develop the regression comparison procedure and to estimate the fairness of the FAST as a predictor of performance in the IERW Program, a sample of the FAST and OAG scores for majority trainees was selected. During the same time period that scores were monitored for the minority samples described in this report, a random sample of 10% of majority officers and 10% of majority WOCs was drawn from the majority population.

The sample sizes for minority/female and majority commissioned officers and WOCs are presented in Table 1.

The Introduction Section of this paper developed the concept that the evaluation of test fairness requires the comparison of minority/female and majority regression lines. A statistical technique was specifically formulated for this purpose by Gulliksen and Wilks (1950). Additionally, there is precedence for the application of this procedure under the mandate of the UGES (Reilly, Zedeck, and Tenopyr, 1979). The Gulliksen Wilks technique, which was derived from Neyman-Pearson likelihood ratio test theory, tests three null hypotheses sequentially (1950, p. 96):

TABLE 1

DESCRIPTIVE STATISTICS FOR THE MINORITY AND MAJORITY COMMISSIONED OFFICER AND WARRANT OFFICER CANDIDATE (WOC) SAMPLES

GROUP	CATEGORY	SAMPLE SIZE	FAST		OVERALL GRADE (OAG)		CORRELATION FAST WITH OAG	SIGNIFICANCE OF THE CORRELATION
			MEAN	STANDARD DEVIATION	MEAN	STANDARD DEVIATION		
BLACK	OFFICER	22	236.36	53.52	85.14	2.64	.537	$\rho<.01$
	WOC	14	331.57	26.11	85.20	2.37	.434	NS
HISPANIC	OFFICER	14	263.57	59.44	85.90	2.40	-.204	NS
	WOC	19	329.16	22.81	84.88	2.13	-.123	NS
ASIAN	OFFICER	3	258.00	46.87	86.10	0.79	.668	NS
	WOC	12	346.00	31.29	86.74	1.64	.079	NS
AMERICAN INDIAN	OFFICER	9	310.67	57.50	85.62	3.29	.337	NS
	WOC	14	336.93	28.44	85.68	1.97	.197	NS
FEMALE	OFFICER	11	257.18	61.03	84.83	4.04	.703	$\rho<.01$
	WOC	18	309.00	8.78	85.74	2.41	.244	NS
MAJORITY	OFFICER	94	293.49	65.14	87.07	2.84	.302	$\rho<.002$
	WOC	117	347.90	29.17	85.83	2.68	.236	$\rho<.01$

1. H1 is the hypothesis that the populations from which the samples were drawn have equal standard errors of the estimate (around the least squares regression line).

2. H2 is the hypothesis that the slopes of the population regression lines are the same.

3. H3 is the hypothesis that the Y-intercepts of the regression lines are equal.

In applying the technique, the three hypotheses are tested sequentially starting with H1. If any hypothesis is rejected, hypothesis testing stops and it is concluded that the samples were drawn from different bivariate populations. If all three null hypotheses are retained, then the samples have the same bivariate dispersion, slope and intercept and thus, coincident regression lines.

In applying the Gulliksen Wilks technique to the current fairness evaluation, a significant problem arises because of the small sample sizes currently available for ethnic and female IERW trainees. Gulliksen and Wilks state that their primary purpose is, ". . . to present large-sample tests for the hypotheses considered from the point of view of Neyman-Pearson likelihood ratio test theory (1950, p. 94)." The smallest sample in the Reilly, et. al. (1979) experiments included 45 subjects. A conservative statistician would prefer to have 100 data points in a "large sample" bivariate distribution. However, it is clear that the sample sizes in the current research, which range from a high of 22 Black Officers to a low of 3 Oriental Officers, do not meet the sample size requirement for the Gulliksen Wilks procedure.

A search of the statistics literature produced a regression line comparison procedure which was derived from the analysis of covariance rather than from Neyman-Pearson likelihood ratio theory. Snedecor and Cochran (1967, pp. 432-436) present a procedure which tests the same three sequential hypotheses discussed by Gulliksen and Wilks (1950). This procedure, while it is sensitive to the usual assumptions made by parametric statistics, is not based on the assumption of large sample sizes.

RESULTS

Table 1 presents sample sizes, means, and standard deviations for the Commissioned Officer and WOC samples. In addition, the correlation of the FAST and overall grade for each group and the significance of that correlation coefficient is shown. At least in part because of the small sample sizes of the minority and female samples, only 2 of the 10 correlations attained significance. In both of the majority samples, the FAST proved to be a significant predictor of overall grade despite the restriction in range caused by the prior use of FAST scores as a selection criterion (Commissioned Officers must score at least 155 and enlisted or civilian entry must score at least 300[2] to gain admission to the IERW training program). In reality, the

[2]Since these data were collected, the FAST cutoff score for WOCs was reduced to 270.

restriction of range problem applies only to the WOC samples since very few of the Commissioned Officer applicants score below 155. The lesser restriction of range in the officer sample is the most probable explanation for the generally higher correlations in that group, as contrasted to the WOC samples.

The three hypotheses tested in the fairness evaluation concern the equality of the standard errors of the estimate, the slopes, and the Y-intercepts for the minority/female and majority regression lines. The logic of the hypothesis test procedure requires that the three hypotheses be tested sequentially. That is, the hypothesis of equal dispersion about the common regression line is tested first. If that F-ratio reaches significance, the hypothesis test procedure stops and it is concluded that the two samples are not taken from the same bivariate population. If the F-test for equality of variance about the common regression line is nonsignificant, then the second hypothesis is tested, i.e., the two slopes are compared. Again, if the F-ratio reaches significance, it is concluded that the two regression lines are not the same. If the F-ratio is nonsignificant, then the third hypothesis is tested, i.e., the Y-intercepts (or elevations) of the two regression lines are compared. Again, if the F-ratio reaches significance, it is concluded that the two samples did not come from the same bivariate population. Only if all three hypothesis tests yield nonsignificant F-ratios can it be concluded that the two regression lines are coincident.

Given the very small population sizes available at the time this research was undertaken, it might be misleading to present hypothesis test results. The statistical power, even in the largest minority/majority comparison, is not sufficiently large to ensure rejection of the null hypotheses if they are indeed false. Thus, these data will be retained and the fairness analysis will be repeated biannually until such time as sufficient data are available to perform a conclusive study.

DISCUSSION

As noted previously, the data base for minority and female IERW trainees is not of sufficient size to permit drawing conclusions regarding the fairness of the FAST as a selection device. The purpose of this paper is to develop the rationale and methodology for such a fairness evaluation. Thus, the current discussion will focus primarily on methodological issues.

In accordance with the UGES the fairness of a selection procedure should be determined by reference to the regression of that selection test (or procedure) on job referenced criteria. Section 14B(3) of the UGES notes that training performance is an acceptable criterion under certain conditions:

> *"Where performance in training is used as a criterion, success in training should be properly measured and the relevance of the training should be shown either through a comparison of the content of the training program with the critical or important work behavior(s) of the job(s), or through a demonstration of the relationship between measures of performance in training and measures of job performance. Measures of relative success in training include but are not limited to instructor evaluations, performance samples, or tests."*

The IERW training program clearly meets the conditions specified in 14B(3) by virtue of the content of the training program and the measures of relative success employed as grading procedures. The curriculum of the IERW Program of Instruction (POI) has been developed specifically to train aviators to perform Army aviation missions in the field. Thus, the content of the training program corresponds very closely to the critical work behaviors performed on the job. Training grades are composed of the three components identified in the UGES: Instructor evaluations (Instructor Pilot put-up scores), performance samples (checkrides), and tests (academic examinations). The IERW overall grade which is used as a criterion in this research is a composite of all three evaluation components. In summary, the design of the current fairness evaluation is in accordance with the directives of the UGES.

While the sample sizes for the minority/female groups presented in Table 1 are too small to justify the drawing of inferences to the entire populations of female and minority aspirant aviators, several points warrant discussion. For both Hispanic samples (Officer and WOC), the FAST has a nonsignificant negative correlation with overall grade. Inspection of the scatter diagrams in both cases reveals that, while the general linear trend is positive for the entire sample, two or three outliers with extreme scores unduly influenced the regression line. For example, in the Commissioned Officer sample, the individual with the highest IERW overall grade, 89.35, has an unusually low FAST score, 197. Expressed as standard scores, this individual's overall grade is $z = 1.44$ whereas his FAST is $z = -1.12$. Conversely, the individual with the lowest overall grade, 79.39, has a moderately high FAST score, 313. Expressed as standard scores, overall grade $z = -2.71$ and FAST $z = .83$. If these two individuals are removed from the distribution, the correlation for the remaining 12 individuals is .193. The sensitivity of this correlation coefficient to only two data points demonstrates the inappropriateness of generalizing from the small minority and female samples in the current study.

The purpose of this research effort is to establish an appropriate methodology to evaluate the FAST for fairness. The methodology reviewed in this paper has been programmed for automated computation on a computer. Additionally, a mechanism has been established to collect data on minority/female and majority IERW trainees. As more minority/female trainees complete pilot training, the fairness evaluation will be iteratively performed until sample sizes permit sufficient statistical power to draw conclusions about the fairness of the FAST.

REFERENCES

Cleary, T. A. *Test bias: Prediction of grades of negro and white students in integrated colleges.* Journal of Educational Measurement, 1968, 5, 115-124.

Cole, N. S. *Bias in selection,* 1972, American College Testing Program, Iowa City, IA.

Cole, N. S. *Bias in selection.* Journal of Educational Measurement, 1973, 10, 237-255.

Cronbach, L. J. *Essentials of psychological testing.* New York: Harper and Brothers, 1960.

Darlington, R. B. *Another look at "culture fairness."* Journal of Educational Measurement, 1971, 8, 71-82.

Guidelines on employee selection procedures. Federal Register, 1970, 35, 149, 12333-12336.

Gulliksen, H. *Theory of mental tests.* New York: John Wiley and Sons, 1950.

Gulliksen, H., and Wilks, S. S. *Regression tests for several samples.* Psychometrika, 1950, 15, 91-114.

Hunter, J. E., Schmidt, F. L., and Rauschenberger, J. M. *Fairness of psychological tests: Implications of four definitions for selection utility and minority hiring.* Journal of Applied Psychology, 1977, 62, 245-260.

Ledvinka, James. *The statistical definition of fairness in the federal selection guidelines and its implications for minority employment.* Personnel Psychology, 1979, 32, 551-562.

McMullen, R. L. *An evaluation of selection rates for minority and female applicants to the US Army initial entry rotary wing flight training program.* US Army Research Institute, Ft Rucker Field Unit, Working Paper, 1981.

Principles for the validation and use of personnel selection procedures, 1975, Division 14 (Division of Industrial and Organizational Psychology), American Psychological Association.

Reilly, R. R., Zedeck, S., and Tenopyr, M. L. *Validity and fairness of physical ability tests for predicting performance in craft jobs.* Journal of Applied Psychology, 1979, 64, 262-274.

Schmidt, F. L., Hunter, J. E., and Urry, V. W. *Statistical power in criterion-related validity studies.* Journal of Applied Psychology, 1976, 61, 473-485.

Snedecor, G., and Cochran, W. G. *Statistical methods* - (6th ed.). Ames: Iowa State University Press, 1967.

Thorndike, R. L. *Concepts of culture fairness.* Journal of Educational Measurement, 1971, 8, 63-70.

Uniform guidelines on employee selection procedures. Federal Register, 1978, 43, 166, August 25, 1978.